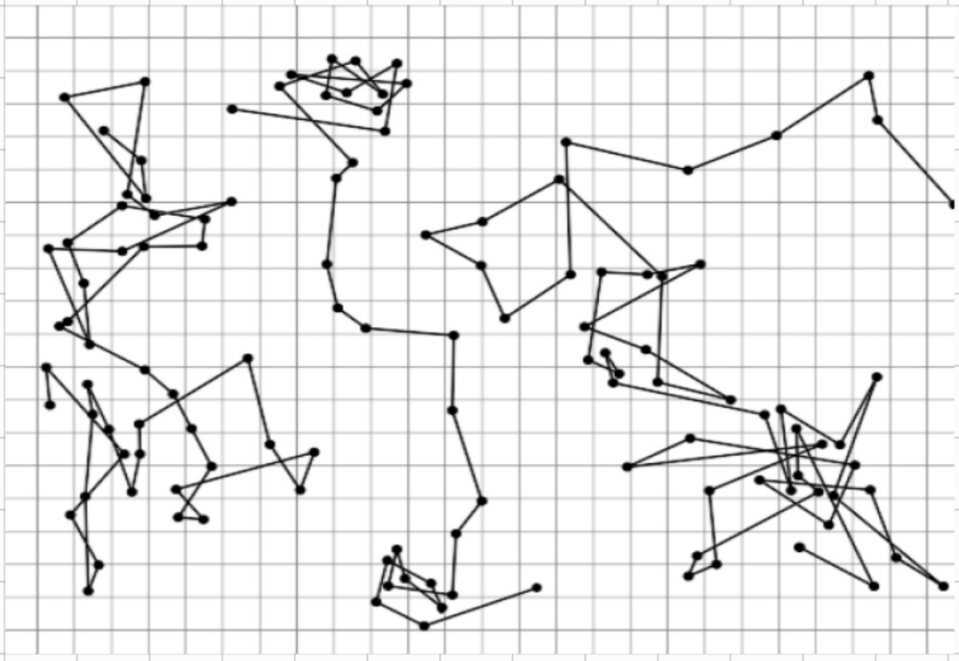
* **Examples** :

1) Does my data come from a prescribe distribution F_X ? "Goodness of fit"

2) Rolling a 6-sided die n times and observe 1, 3, 1, 6, 4, ... Is it a fair die?

3) Einstein's Theory of Brownian Motion



* Einstine's Theory of Brownian Motion

P_t : position of the particular a time t.

$P_{t+\Delta t}$: position of the particular at time $t+\Delta t$

$$P_{t+\Delta t} \sim N\left(P_t, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right) \text{ where } \sigma^2 = \frac{RT(\Delta t)}{3\pi \eta r N_A}$$

We want to test the goodness of fit.

Experiment : Measured the position of a particle every 30s

to verify Einstein's theory and computed N_A.
He figured out $a^2 = 2.23 \times 10^{-7}\ cm^2$

Question : Does Perrin's data fit with Einstein's model ? ToH - Question.

* Def : Hypothesis Test

It is a Binary question about the data distribution. Our goal is to either accept a null hypothesis (H_0 : specifies something about the distribution) or to reject null hypothesis H_0 in favour of an alternative hypothesis H_1.

If H_0 (similarly H_1) completely specifies the probability distribution for the data then, it is called a simple hypothesis. Otherwise, it is called composite hypothesis.

* Rolled a die n times and observe the outcome.
X_i denotes the number of time we obtain i, $i \in \{1, \dots, 6\}$

$H_0 : X_i \sim \text{Multinomial}\left(n, \left(\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right)\right)$
$H_1 : X_i \sim \text{Multinomial}\left(n, \left(\frac{1}{9}, \frac{1}{9}, \frac{2}{9}, \frac{2}{9}, \frac{1}{9}, \frac{2}{9}\right)\right)$
Both are simple hypothesis.

Composite alternative hypothesis :
$H_1 : X_i \sim \text{Multinomial}(n, (p_1, \dots, p_6))$; $p_i \neq \frac{1}{6}$ $\forall i$

* Another problem is creating a Test Statistic.

$T := T(X_1, \ldots, X_n)$ is any statistic; extreme values (large / small) of T provide evidence against H_0.

$$T = \left(\frac{X_1}{n} - \frac{1}{6}\right)^2 + \cdots + \left(\frac{X_6}{n} - \frac{1}{6}\right)^2.$$

Large of T provide evidence against the null hypothesis of a fair die.

Going Back to Brownian Motion:

Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be the displacement vector $P_{30} - P_0$, $P_{60} - P_{30}$, ... where $P_t \in \mathbb{R}^2$ is the position of the particle at time t in perrin's experiment.

Simple: $H_0: (X_i, Y_i) \overset{iid}{\sim} N(0, 2.23 \times 10^{-7} I)$

Composite: $H_1: (X_i, Y_i) \overset{iid}{\sim} N(0, v^2 I)$ for some $v^2 > 0$

Test Statistic: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

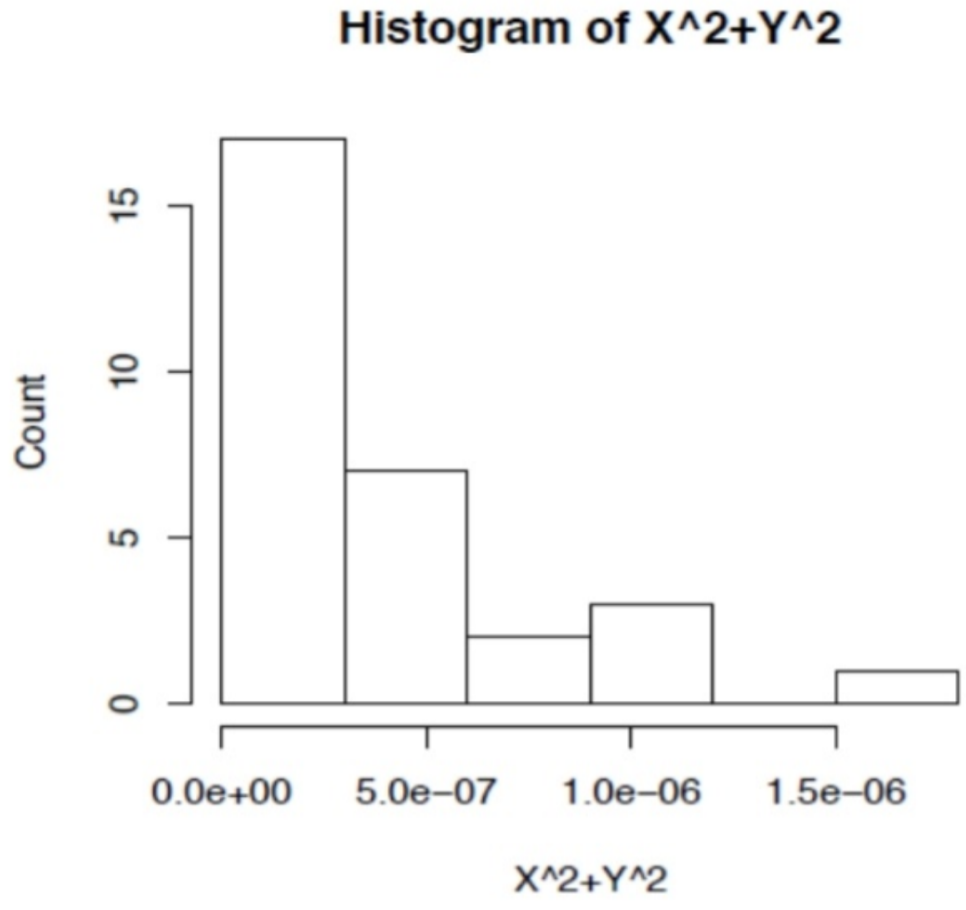
$$V = \frac{1}{n} \sum_{i=1}^{n} (X_i^2 + Y_i^2)$$

Large / smaller values than 0 for $\bar{X}$ and $\bar{Y}$ provide evidence against H_0. Values larger / smaller than 2.23×10^{-7} provide evidence against H_0.

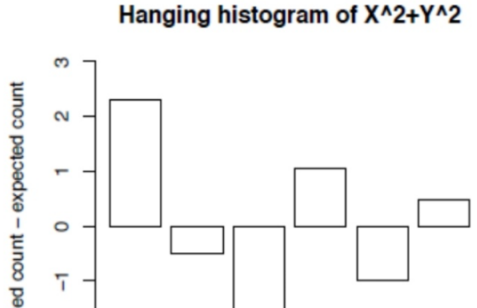
We can rewrite: $V = \sum_{i=1}^{n} R_i$, $R_i = X_i^2 + Y_i^2$ where $R_i \sim 2.23 \times 10^{-7} \chi^2_2$ (as we have two variables)

* How to derive test statistic from hypothesis ;

Let $R_i = X_i^2 + Y_i^2$. We are interested in testing $R_1, \ldots, R_n$ are distributed as $2.23 \times 10^{-7}\ \chi^2_2$ distribution (under H_0).



Deviations from σ_0^2 are better visualized by a hanging histogram which plots $O_i - E_i$ where O_i is the observed count for bin i and E_i is the expected count for bin i under the H_0 distribution.



A test statistic can be $T = \sum_{i=1}^{6} (O_i - E_i)^2$.

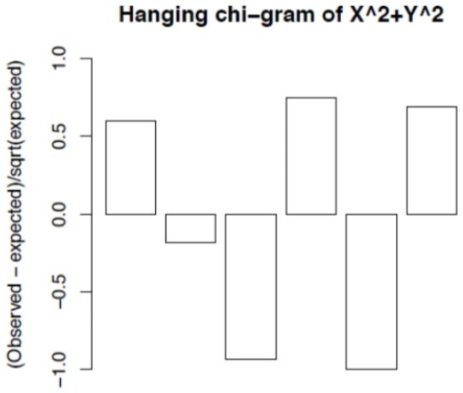
Let p_i be the probability that the hypothesized χ^2 dist. assigns to bin i.

If H_0 is true then $O_i \sim Bin(n, p_i)$ so $E_i = \mathbb{E}(O_i) = np_i$ and $V(O_i) = np_i(1-p_i) = \mathbb{E}(O_i - E_i)^2$.

The variation in O_i is smaller and it scales approximately linearly with p_i if p_i is close to 0.

We want to stabilize the variance : $\frac{O_i - E_i}{\sqrt{E_i}} = \frac{O_i - E_i}{\sqrt{np_i}}$

hanging chi-gram plot



The test statistic $T = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$ is called **Pearson's chi-squared statistic for goodness of fit.**[3]

* Test statistic for Q-Q plots :

A Q-Q plot (or probability plot) compares the sorted

values of R_i with $\frac{1}{n+1}, \frac{2}{n+1}, \ldots, \frac{n}{n+1}$ quantiles of the hypothesized $\sigma_0^2 \chi^2_2$ distribution.

Values close to the line $y = x$ indicate a good fit. F is the of $\sigma_0^2 \chi^2_2$ distribution $|R_{(i)} - F^{-1}(\frac{i}{n+1})|$.

One way is to take the maximum vertical deviation from the $y = x$ line.

Let $R_{(1)}, \ldots, R_{(n)}$ be the ordered statistic of $R_1, \ldots, R_n$. Then, we consider $T = \max_{1 \le i \le n} |R_{(i)} - F^{-1}(\frac{i}{n+1})|$ where $F^{-1}(t)$ is the t^{th} quantile.

For the values of R where the distribution has high density, the quantiles are closer together so we expect a smaller vertical deviation. This explain why we see more vertical deviations in the upper right of the Q-Q plot.

Another test statistic: $T = \max_{1 \le i \le n} |F(R_i) - \frac{i}{n+1}|$

One sample Kolmogrov Smirnov (K-S) statistic.
$T_{KS} = \max_{1 \le i \le n} (\max\{|F(R_i) - \frac{i-1}{n}|, |F(R_i) - \frac{i}{n}|\})$.

For large n: $T_{KS} \longrightarrow T$

* Null Hypothesis:

	Ho True	Ho False
Ho reject	Type 1	Correct
Ho not reject	Correct	Type 2

α - level of significance.

* Null Distribution & Type 1 error :

Let us consider our statistic T, how large (or small) T needs to be before we can safely assert that H_0 is false.

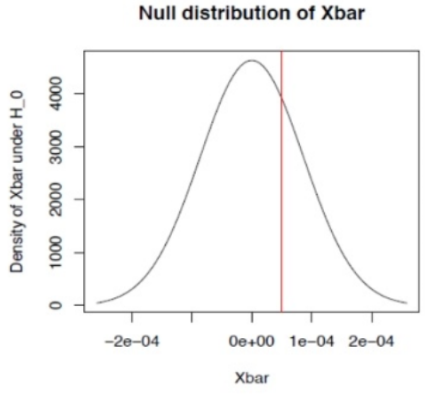
In most cases, we can never be 100% sure that H_0 is false but we can compute T from the observed data and compare it with the sampling distribution of T if H_0 were true. This is called the null distribution of T.

Under H_0 : $\bar{X} \sim N(0, \frac{\sigma_0^2}{n})$ and $\bar{Y} \sim N(0, \frac{\sigma_0^2}{n})$
as $(X_i, Y_i) \overset{iid}{\sim} N(0, \sigma_0^2 I)$

Based on the observed data $\bar{x} = 0.5 \times 10^{-4}$ or another data ; $\bar{x} = 2.5 \times 10^{-4}$

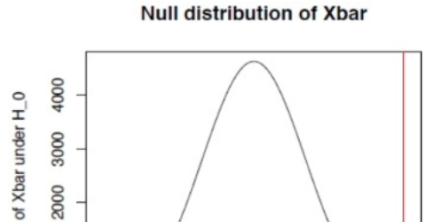
Under $H_0, \bar{X} \sim \mathcal{N}\left(0, 2.23 \times 10^{-7}/n\right)$. This normal distribution is the null distribution of $\bar{X}$.

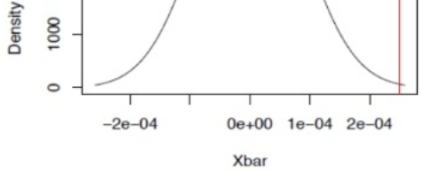
Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:



If, for the observed data, $\bar{X} = 0.5 \times 10^{-4}$, this would not provide strong evidence against H_0. In this case, we might accept H_0.

Here's the PDF for the null distribution of $\bar{X}$, when $n = 30$:





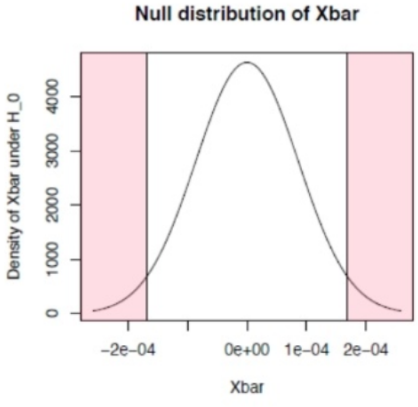
If, for the observed data, $\bar{X} = 2.5 \times 10^{-4}$, this would provide strong evidence against H_0. In this case we might reject H_0.

The rejection region is the set of values of T for which we would reject H_0. The acceptance region is the set of values of T for which we choose to accept H_0.
$\mathbb{P}_{H_0}(\text{reject } H_0) = \alpha$, the value α is the significance level.

Under H_0, T belongs to the rejection region with probability α then the test is level-α test.

Two sided level-α test reject H when T falls in the shaded region.

(Notation: For a simple null hypothesis H_0, we write $\mathbb{P}_{H_0}[\mathcal{E}]$ to denote the probability of event $\mathcal{E}$ under H_0, i.e., the probability of $\mathcal{E}$ if H_0 were true.)



Example 1.6.2. *A (two-sided) level-α test might reject H_0 when $\bar{X}$ falls in the above shaded regions. Mathematically, let z_α denote the $1-\alpha$ quantile, or "upper α point", of the distribution $\mathcal{N}(0,1)$. As $\bar{X} \sim \mathcal{N}\left(0, \sigma^2/n\right)$ under H_0 (where $\sigma^2 = 2.23 \times 10^{-7}$), the rejection region should be $\left(-\infty, -\frac{\sigma}{\sqrt{n}} \times z_{\alpha/2}\right] \cup \left[\frac{\sigma}{\sqrt{n}} \times z_{\alpha/2}, \infty\right)$.*

Acceptance region : $\left]-\frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\right[$

Rejection region : $]-\infty, -\frac{\sigma}{\sqrt{n}} z_{\alpha/2}] \cup [\frac{\sigma}{\sqrt{n}} z_{\alpha/2}, +\infty[$

* **Def** : P - Value

The smallest significance level at which your test would be rejecting H_0.

Let the test be a one sided test that reject for large values of T if t_{obs} denote the value of T computed based on the obs data $\quad p\text{-value} = \mathbb{P}_{H_0}(T \geq t_{obs})$

Two sided: $\quad p\text{-value} = 2 \min(\mathbb{P}_{H_0}(T \geq t_{obs}), \mathbb{P}(T \leq t_{obs}))$

* p-value is a quantitive measure of the extent to which the data supports (or does not support) H_0.

* Failing to reject $H_0 \not\Rightarrow$ there is strong evidence that H_0 is true

A) The particular test statistic chosen is not good at distinguishing the null hypothesis from the true distribution. Equivalently, the true distribution is not well captured by H_1 that your statistic is capturing.

B) You don't have enough data to reject H_0 at the desired significance level. In this case, your study is underpowered.

* Suppose that $H_0 : X_1, \dots, X_n \sim N(8, 1)$ and $T = \bar{X}$ so the exact null distribution is $N(8, 1/n)$.

* Determining the null distribution:

1) Exact Null Distribution
2) Derive an asymptotic approximation, using CLT & continuous mapping theorem (when H_0 is simple, we can obtain null distribution by simulation).

* Deriving asymptotic null distribution:

Let $X_1, \dots, X_6$ denote the counts of $1, 2, \dots 6$ from n rolls of a die, and consider testing if the die is fair.

H_0: $(X_1, \dots, X_6) \sim \text{Multinomial}\left(n, \left(\frac{1}{6}, \frac{1}{6}, \dots, \frac{1}{6}\right)\right)$

Test statistic: $T = \sum_{i=1}^{6} \left(\frac{X_i}{n} - \frac{1}{6}\right)^2$

$X_i \sim \text{Bin}\left(n, \frac{1}{6}\right)$ by CLT $\frac{X_1 - n/6}{n \cdot 1/6 \cdot 5/6} \sim N(0,1)$ and as $n \to +\infty$ $\frac{X_1 - n/6}{n} = \frac{X_i}{n} - \frac{1}{6} \sim N(0,1)$

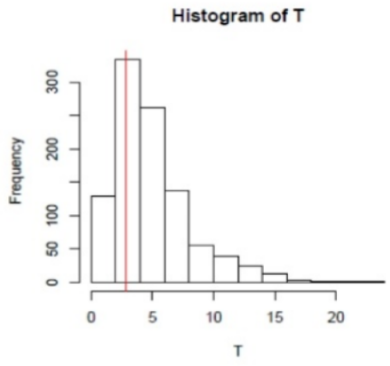
Hence, $\left(\frac{X_i}{n} - \frac{1}{6}\right) \sim X_1^2 \Rightarrow T \sim X_5^2$ as $(X_1, \dots, X_6)$ are not jointly independent

Thus, T approximately distributed as $\frac{1}{6n} X_5^2$ distribution

Rejection region is if $T > c$, c is derived from the level of significance. We get c from the $1-\alpha$ quantile of $\frac{1}{6n} X_5^2$ so $c = \frac{1}{6n} X_5^2(\alpha)$

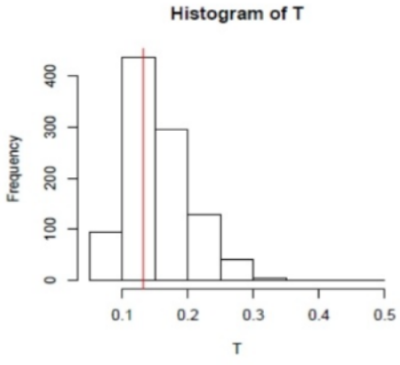
1.11 Using a simulated null distribution

Example 1.11.1. *Let T be Pearson's chi-squared statistic for goodness of fit for the values $X_1^2 + Y_1^2, \dots, X_{30}^2 + Y_{30}^2$ from Perrin's experiments. We may simulate the null distribution of T :*



This shows the 1000 values of T across 1000 simulations. The observed value $t_{\text{obs}} = 2.83$ for Perrin's real data is in red.

Example 1.11.2. *Let T be the K-S statistic for $X_1^2 + Y_1^2, \ldots, X_{30}^2 + Y_{30}^2$. We may simulate the null distribution of T :*



The observed value $t_{\text{obs}} = 0.132$ for Perrin's real data is in red.

We obtain an approximate p-value as the fraction of simulated values of T larger than t_{obs}. (For a two-sided test, we would take either the fraction of simulated values of T larger than t_{obs} or smaller than t_{obs}, and multiply this by 2.)

For Perrin's data, the Pearson chi-squared p-value is 0.754, and the K-S p-value is 0.612. We accept H_0 in both cases, and neither test provides significant evidence against Einstein's theory of Brownian motion.

* Simple alternative and the Neyman Pearson Lemma

- Till now, we discussed a number of ways to construct test statistics for testing a simple null hypothesis, and we showed how to use the null distribution of the statistic to determine the rejection region so as to achieve the desired significance level.
- In this section, our goal is to answer the following question: Which test statistic

should we use?

- The answer depends on the alternative hypothesis that we wish to distinguish from the null.

We will focus on simple H_0 against simple H_1.

$\beta = \mathbb{P}_{H_1}(\text{Accept } H_0)$ probability of type 2 error

$1-\beta = \mathbb{P}_{H_1}(\text{Reject } H_0)$ probability of correctly rejecting H_0 called the power of the test.

Goal: Maximize the power of the test subject to the significance level of the test under H_0 is atmost α

This is a constrained optimization problem

Let X be the data vector, $X = (X_1, \ldots, X_n)$ and let $x = (x_1, \ldots, x_n)$ be vector of possible values of X.

H_0: X is distributed with joint PMF $f_0(x) := f_0(x_1, \ldots, x_n)$

H_1: X is distributed with joint PMF $f_1(x) := f_1(x_1, \ldots, x_n)$

Let $\mathcal{X}$ be the set of all both values of X under f_0 & f_1. We need to specify the rejection region $R \subset \mathcal{X}$ such that we reject H_0 if the observed data belongs to R otherwise we accept H_0.

$$\mathbb{P}_{H_0}(\text{Reject } H_0) = \alpha = \sum_{x \in R} f_0(x)$$

$$\mathbb{P}_{H_1}(\text{Reject } H_0) = 1-\beta = \sum_{x \in R} f_1(x)$$

Formally, we define the optimization problem as choosing the rejection region $R \subset X$ with the goal maximize $\sum_{n \in R} f_1(n)$ subject to $\sum_{n \in R} f_0(n) \leq \alpha$.

We want to know what are the best points n to include in the rejection region R. Hence, R should consist of those points n corresponding to the smallest values of $\frac{f_0(n)}{f_1(n)}$, as those give the smallest increase in type 1 error for unit increase in the power. These points give most significant evidence against H_0 and in favour of H_1 over H_0. The statistic $L(x) = \frac{f_0(x)}{f_1(x)}$ is the likelihood ratio statistic and it will reject for small values of $L(x)$.

Example 2.2.1. *Consider data $X_1, \ldots, X_n$ and the following null and alternative hypotheses:*

$$H_0 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(0, 1)$$
$$H_1 : X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, 1).$$

Here we assume μ is a known, specified value (not equal to 0), so that H_1 is a simple alternative hypothesis. The joint PDF *of $(X_1, \ldots, X_n)$ under H_0 is*

$$f_0(n_1, \ldots, n_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-n_i^2/2} = (2\pi)^{-n/2} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} n_i^2}$$

$$f_1(n_1, \ldots, n_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(n_i-\mu)^2}{2}} = (2\pi)^{-n/2} \cdot e^{-\frac{1}{2}\sum_{i=1}^{n} (n_i-\mu)^2}$$

$$L(n_1, \ldots, n) = \frac{f_0(n_1, \ldots, n_n)}{f_1(n_1, \ldots, n_n)} = \exp\left\{ \frac{-1}{2} \sum_{i=1}^{n} n_i^2 + \frac{1}{2} \sum_{i=1}^{n} (n_i - \mu)^2 \right\}$$

$$= \exp\left\{ \frac{1}{2} \sum_{i=1}^{n} \mu^2 - \mu \cdot \sum_{i=1}^{n} n_i \right\}$$

$$= \exp\left\{ \frac{n}{2} \mu^2 - \mu n \bar{n} \right\}$$

$$= \exp\left\{ \frac{n}{\alpha} \mu \left(\mu - 2\bar{x} \right) \right\}$$

$$= \exp\left\{ \frac{n\mu^2 - 2\mu\bar{x}}{\alpha} \right\}$$

If $\mu > 0$, then $L(x)$ is a decreasing function of $\bar{x}$. Hence, rejecting for small values of $L(x)$ is equivalent to rejecting large values of $\bar{x}$.

Rejecting H_0 if $\bar{X} > c$. Under H_0, $\bar{X} \sim N(0, 1/n)$, $c = \frac{1}{\sqrt{n}} z_\alpha$ where z_α is the upper α point of the $N(0,1)$.

If $\mu \leq 0$, then $L(x)$ is an increasing function of $\bar{x}$. Hence, rejecting for small values of $L(x)$ is equivalent to rejecting small values of $\bar{X}$.

Rejecting H_0 if $\bar{X} < c$. Under H_0, $\bar{X} \sim N(0, 1/n)$, $c = \frac{-1}{\sqrt{n}} z_\alpha$ where $-z_\alpha$ is the upper $1-\alpha$ point of the $N(0,1)$. $\bar{X} < \frac{-z_\alpha}{\sqrt{n}}$

If we change it to $H_1 : X_1, \ldots, X_n \sim N(8, 1)$ and $H_1 : X_1, \ldots, X_n \sim N(10^3, 1)$ as we are not considering the value of α then we only consider the case $\mu > 0$.

In this case as $\mu \neq 0$ so we have two powerful tests one of each side so we do not have one uniformly powerful test $\forall n$

* **Lemma** : Neyman Pearson

Let H_0 and H_1 be simple hypotheses, for a constant $c > 0$, suppose that the likelihood ratio test rejects H_0 when $L(x) < c$ has significance level α. Then, for any other test of H_0 with atmost significance level α, its power

against H_1 is at most the power of the likelihood ratio test.

* **Proof** :

Consider the discrete case and let $R = \{x \mid L(x) < \eta\}$ be the rejection region of the Likelihood Ratio Test. Note that among all the subsets of $\mathcal{X}$, R maximize the quantity $\sum_{x \in R} (c f_1(x) - f_0(x))$ because $c f_1(x) - f_0(x) > 0 \ \forall x \in R$ and $c f_1(x) - f_0(x) \leq 0 \ \forall x \notin R$.

Let R' be another rejection region. Then,

$$\sum_{x \in R'} (c f_1(x) - f_0(x)) \leq \sum_{x \in R} (c f_1(x) - f_0(x))$$

$$\Rightarrow c \left(\sum_{x \in R} f_1(x) - \sum_{x \in R'} f_1(x) \right) \geq \sum_{x \in R} f_0(x) - \sum_{x \in R'} f_0(x)$$

$$= \alpha - \sum_{x \in R'} f_0(x) \leq \alpha$$

$$\geq 0$$

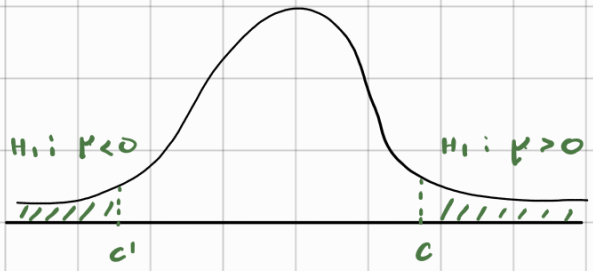
Hence, $\sum_{x \in R} f_1(x) \geq \sum_{x \in R'} f_1(x)$ so we deduce the power of the LRT is at least that of the other test. Similarly, for the continuous case.

* **Uniformly Most Powerful Test** :

The most powerful test against the alternative $H_1 : X_1, \dots, X_n \sim N(\mu, 1)$ is the same for $\mu > 0$ (reject if $\bar{X} > \frac{1}{\sqrt{n}} z_\alpha$) and neither the test the test statistic nor the rejection region depends on the specific value of μ. This means that, in fact, this test is uniformly most powerful (UMP) against the one sided composite hypothesis.

$H_1: X_i \sim N(\mu, 1)$ $\forall i = 1, \dots, n$ where $\mu > 0$ or $X_i \sim N(\mu, 1)$ where $\mu < 0$

On the other hand, the MP test is different for $\mu > 0$ vs. $\mu < 0$, one rejects for large positive values of $\bar{x}$ and the other rejects for large negative values of $\bar{x}$. Hence, a single MP test does not exist for the two sided hypothesis $H_1: X_i \sim N(\mu, 1)$ $\forall i = 1, \dots, n$, $\mu \neq 0$



$H: \mu \neq 0$ both rejection regions

* **Example** :

Find the MP test at level of significance α to test whether the data $X_1, \dots, X_n \sim \text{Ber}(p)$ where $H_0: p = \frac{1}{2}$ $H_1: p \neq \frac{1}{2}$

$$L(X) = \frac{f_0(x_1, \dots, x_n)}{f_1(x_1, \dots, x_n)} = \frac{\prod_{i=1}^{n} (\frac{1}{2})^{x_i} (\frac{1}{2})^{1-x_i}}{\prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}} = \frac{1}{\alpha^n} \frac{1}{p^{\sum x_i} (1-p)^{n - \sum x_i}}$$

$$L(X) = \frac{1}{\alpha^n (1-p)^n} \left(\frac{1-p}{p} \right)^{\sum_{i=1}^{n} x_i}$$

$p < \frac{1}{2}$: Hence, $L(X)$ is an increasing function of $S = \sum_{i=1}^{n} X_i$ so we reject for small values of S. By NP lemma, the most powerful test rejects H_0 when $S < c$ for some c.

$H_1: X_1, \dots, X_n \sim \text{Bernoulli}(p)$, $p < \frac{1}{2}$

$S \sim \text{Binomial}(n, \frac{1}{2})$ under H_0 so c is the α quantile of $\text{Binomial}(n, \frac{1}{2})$.

$p > \frac{1}{2}$: Hence, $L(x)$ is a decreasing function of $S = \sum_{i=1}^{n} X_i$ so we reject for large values of S. By NP lemma, the most powerful test rejects H_0 when $S > c$ for some c.

H_1 : $X_1, \ldots, X_n \sim$ Bernoulli (p) , $p > \frac{1}{2}$

$S \sim$ Binomial $(n, \frac{1}{2})$ under H_0 so c is the $(1-\alpha)$ quantile of Binomial $(n, \frac{1}{2})$.

This test is the same for all $p < \frac{1}{2}$ $(p > \frac{1}{2})$ and it is in fact the uniformly most powerful test against the one sided composite hypothesis H_1 : $X_1, \ldots X_n \sim$ Ber(p) $(p < \frac{1}{2}, p > \frac{1}{2})$.

* **Remark** :

We have glossed over a detail which is that when the distribution of $L(X)$ is discrete under H_0, it might not be possible to choose c ; the significance level exactly α.

Example: $S \sim$ Bin$(20, \frac{1}{2})$

$\mathbb{P}(S \geq 15) = 0.021$ $\mathbb{P}(S \geq 14) = 0.058$

If we reject H_0 when $S \geq 14$, we don't reach the significance level α. If we reject H_0 when $S \geq 15$, we are too conservative.

Theoretically, we use a randomized test. Always reject H_0 when $S \geq 15$, we accept H_0 when $S \leq 13$ and reject H_0 with

a certain probability when $S = 14$ where this probability is chosen to make the significance level is exactly α. By NP lemma, this is the most powerful test among all randomized tests. In practice, it might not be acceptable to use randomized test, so we might take the more conservative option when $S \geq 15$.

* Hypothesis
  - Simple Hypothesis (H_0, H_1)
  - Composite Hypothesis
    - Two sided $(\theta \neq \theta_0)$
    - One sided
      - $\theta > \theta_0$
      - $\theta < \theta_0$

* **Example** : Suppose that there are 80 students in a class taking MATH 350 course. A diagnistic example is administered at the start of the quarter, and a comparable exam is administered at the end of the quarter. Did the MATH 350 course improve students knowledge of statistics?

Let X_i be the difference of scores of the two tests for the i^{th} student. Assume $X_1, \dots, X_{80} \sim N(\mu, \sigma^2)$.

We might formulate the hypothesis that $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$. Both hypotheses are composite as σ^2 is unknown.

If we cannot make the normality assumption then we might assume that $X_1, \dots, X_{80}$ are i.i.d with some pdf f and then we test.

H_0 : f is symmetric around zero

H_1 : f is symmetric around μ

We might drop the symmetry assumption of the test. Then,

H_0 : f has median zero

H_1 : f has median μ ($\mu > 0$)

Selection of the test : 1) Some prior knowledge on the distribution of the scores.

2) Visual Implementation

It's hard to make two exams that are equally difficult. What if the second exam happened to be easier ?

To address this we add a control group. We give 100 other students who are not taking MATH 350 course the same two exams at the start & at the end of the quarter. Let Y_i be the difference in scores for the student i for the control group. If we believe that $Y_1, Y_2, \dots, Y_{100} \overset{iid}{\sim} N(\mu_Y, \sigma^2)$ and $X_1, \dots, X_8 \overset{iid}{\sim} N(\mu_X, \sigma^2)$.

To do this, we might formulate the test as $H_0 : \mu_X = \mu_Y$

$H_1 : \mu_X > \mu_Y$

If we cannot have the normality assumption then we might test

H_0 : $f = g$ where f is the PDF of the X_i's

H_1 : f stochastically dominates g and g is the PDF of the Y_i's.

We can rewrite the alternative hypothesis as $\mathbb{P}(X \geq x) \geq \mathbb{P}(Y \geq y)$ $\forall x \in \mathbb{R}$.

We have to inspect the data again to check the normality assumption.

* Computing the value of α :

$X_1, \dots, X_n \sim N(\mu, \sigma_0^2)$ Test statistic : $\bar{X}$

Reject H_0 : $\bar{X} > c$

H_0 : $\mu = 80$

H_1 : $\mu > 80$

Under H_0 : $\bar{X} \sim N(80, \frac{\sigma_0^2}{n})$

$$\alpha = \mathbb{P}_{H_0}(\bar{X} > c) \quad , \quad \alpha = 0.05$$

$$= \mathbb{P}\left(\frac{\sqrt{n}(\bar{X} - 80)}{\sigma_0} > \frac{\sqrt{n}(c-80)}{\sigma_0}\right)$$

$$= \mathbb{P}(Z \geq c^*)$$

$$0.95 = \mathbb{P}(Z \leq c^*)$$

$$\Rightarrow 1.65 = \frac{\sqrt{n}(c-80)}{\sigma_0} \quad , \quad \sigma_0 = 10, \; n = 100$$

$$\Rightarrow c = 81.65$$

Rejecting H_0 when $\bar{X} > 81.65$ with 0.05 significance level.

Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$; $H_0: \mu = 80$ $H_1: \mu > 80$ where σ^2 is unknown. Composite H_0 & H_1

What is the distribution of the test statistic $\bar{X}$? Calculate Type I error.

When testing a composite H_0 against a composite H_1. When testing a composite H_1, there is a probability of Type I error associated to each of the data distributions $P \in H_0$ (the prob of rejecting H_0 if the true distribution is P) and a probability of type II error associated to each of the data distributions $P \in H_1$ (prob of accepting H_0 when P is the true distribution). A test has significance level α if the maximum probability of Type I error for any $P \in H_0$ is α. This means that to design a level α-test, we need to control the prob. of type I error for every $P \in H_0$. This is difficult !

* One sample t - test :

Assume $X_1, X_2, \dots, X_n \sim N(\mu, \sigma^2)$ for unknown μ and σ^2. Consider testing :

$H_0 : \mu = 0$

$H_1 : \mu > 0$

σ^2 is unknown, a natural idea is to estimate σ^2 by the sample variance : $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ and consider the statistic :

$$T = \frac{\sqrt{n}(\bar{x} - \mu_0)}{S} = \frac{\sqrt{n}(\bar{x}/\sigma)}{S/\sigma} \sim \frac{N(0,1)}{\sqrt{\frac{1}{n-1}\chi^2_{n-1}}} \equiv t_{n-1}$$, N, χ^2 are independent

The distribution of t does not depend on σ, so it is same for any $P \in H_0$.

* **Def** : t - distribution

If $Z \sim N(0,1)$ and $v \sim \chi^2_n$; Z and v are independent then the distribution of $\frac{Z}{\sqrt{\frac{1}{n} v}}$ is called the t-distribution with n df denoted by t_n.

Under H_0, $T \sim t_{n-1}$. Let $t_{n-1,\alpha}$ denote the upper α point of the distribution t_{n-1}, then the test that rejects for $T > t_{n-1,\alpha}$ is called the one sample t - test.

The one sample t - test is often used in paired two sample setting ; the previous example. There we actually have two paired samples before and after the test per student. We perform the test by first considering the differences of the

scores, such settings are called paired two sample t-test although we perform a one sample t-test.

* Two sample t-test :

Let two independent samples $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_x, \sigma^2)$ and $Y_1, \dots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma^2)$. Suppose that they have a common variance.

$H_0 : \quad \mu_x - \mu_y = 0$

$H_1 : \quad \mu_x - \mu_y > 0$

A natural idea is to reject H_0 for large values of $\bar{X} - \bar{Y}$. $\bar{X} \sim N(\mu_x, \frac{\sigma^2}{n})$

$\bar{Y} \sim N(\mu_y, \frac{\sigma^2}{m}) \Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$

Under H_0 : $T \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right) \Rightarrow \dfrac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2(1/n + 1/m)}} \sim N(0,1)$

- If σ^2 is known : $\dfrac{\bar{X} - \bar{Y}}{\sqrt{\sigma^2(1/n + 1/m)}} > z_\alpha$ then we reject H_0

- If σ^2 is unknown : $S_p^2 = \dfrac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2\right)$
  is called the pooled sample variance and it is used to estimate σ^2.

$$\frac{S_p^2}{\sigma^2} \sim \frac{1}{m+n-2}\chi^2_{m+n-2}$$

$$T = \frac{(\bar{X} - \bar{Y}) / \sqrt{\sigma^2(1/n + 1/m)}}{\sqrt{S_p^2/\sigma^2}} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(1/n + 1/m)}} \sim t_{n+m-2}$$

$$\mathbb{P}_{H_0}(T > c) = \alpha \Rightarrow \mathbb{P}(T > t_{n+m-2,\alpha}) = \alpha$$

$\Rightarrow \quad c = t_{n+m-2,\alpha}$

We reject H_0 when $T > t_{n+m-2,\alpha}$ (upper α ($1-\alpha$) point of t_{n+m-2}) for the 2 sample t-test.

* **Remark** : The assumption of common variance σ^2 for the samples is often violated in practice. We can assume that $X_1, \ldots, X_n \sim N(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_n \sim N(\mu_y, \sigma_y^2)$ for $\sigma_x^2 \neq \sigma_y^2$.

$Var(\bar{X} - \bar{Y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$, we may estimate this by $\frac{s_x^2}{n} + \frac{s_y^2}{m}$ where s_x^2, s_y^2 are the sample variances.

Then, we may use the test statistic :

$$T_{Welch} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}}$$

T_{Welch} under H_0 is not exactly t-distribution but it was shown by Welch to be close to a t-distribution with

$$\frac{\left(s_x^2/n + s_y^2/m\right)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}} \text{ df.}$$

The test that rejects H_0 when T_{Welch} exceeds the upper α point of this distribution is called the Welch t-test or the unequal variances t-test.

* Testing of Proportions :

Lets assume you roll a 4-faced die with values $\{1, 2, 3, 4\}$ 100 times and the value "4" appears 29 times, does this data suggest your die is unbaised.

If the die is unbaised we can expect the proportion of "4" to be 0.25. Let the proportion of "4" be p.

$H_0 : p_0 = 0.25$ vs $H_1 : p_1 \neq 0.25$

We observe the proportion of 4 to be $\hat{p} = 0.29$.

Let Y be the number of times "4" appears Under H_0:

$\mathbb{E}(Y) = n\hat{p} = np_0$

$Var(Y) = n\hat{p}(1-p) = np_0(1-p_0)$

$\mathbb{E}(\hat{p}) = \mathbb{E}\left(\frac{Y}{n}\right) = \frac{1}{n}\mathbb{E}(Y) = p_0$

$Var(\hat{p}) = Var\left(\frac{Y}{n}\right) = \frac{1}{n^2}Var(Y) = \frac{p_0(1-p_0)}{n}$

By CLT : $\hat{p} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$ as $n \to +\infty$

$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ → observed

Critical : $z_{\alpha/2}$

Acceptence Region : $]-z_{\alpha/2}, z_{\alpha/2}[$

* Testing of Proportion:

Times magazine reported the results of a telephone poll of 800 adults. The question was "should America increase the federal tax on cigarretes to pay for health care reforms?". The results are as follows:

	Non-Smokers (1)	Smokers (2)

n_i	# of participants	605	195
y_i	# of "yes" response	351	41

Can you conclude that the 2 groups of individuals significantly differ in their opinion ?

Let p_1, p_2 denote the proportion of individuals responding "yes" in the two subgroups.

$$\hat{p}_1 = \frac{y_1}{n_1} = \frac{351}{605} = 0.58$$

$$\hat{p}_2 = \frac{y_2}{n_2} = \frac{41}{195} = 0.21$$

$H_0 : p_1 = p_2$

$H_1 : p_1 \neq p_2$

By CLT: $\hat{p}_1 \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right)$

$$\hat{p}_2 \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

$$\hat{p}_1 - \hat{p}_2 \sim N\left(0, p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right), \quad \text{Pooled Estimate of } p: \hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

$$z_{obs} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_{\alpha/2} \text{ or } z_{obs} < -z_{\alpha/2} \text{ reject.}$$

* The theoritical work on the hypothesis testing of variance is primarily based on fitting a Normal distribution to the data & then the use of sampling distribution.

* One Sample Test :

Given $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. We aim to test whether

H_0: $\sigma^2 = \sigma_0^2$ against the alternatives

H_1: $\sigma^2 \neq \sigma_0^2$, $\sigma^2 > \sigma_0^2$, $\sigma^2 < \sigma_0^2$

$$S^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1} \qquad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Under H_0: $$\frac{(n-1) S^2}{\sigma_0^2} \sim \chi^2_{n-1}$$

Rejection Region:

$\sigma^2 \neq \sigma_0^2$

$\chi^2 > \chi^2_{\alpha/2, n-1}$ or $\chi^2 < \chi^2_{1-\alpha/2, n-1}$

$\sigma^2 > \sigma_0^2$

$\chi^2 > \chi^2_{\alpha, n-1}$

$\sigma^2 < \sigma_0^2$

$\chi^2 < \chi^2_{1-\alpha, n-1}$

* A manufacturer of hazel safety hats is concerned about the mean and variation of the force its helmets transmits to wearers when subjected to external force. They designed the helmets so that the mean force transmitted is 800 powels with s.d to be less than 40 pounds. Tests on 40 helmets were run and $\bar{x}$ = 825 pounds and S = 48.5 pounds. Do the data provide sufficient evidence that the population sd exceeds 40 pounds. Do the data provide sufficient evidence at $\alpha = 0.05$ to conclude that sd $\neq$ 40 pounds.

$\chi^2_{0.05, 39} = 54.572$ $\qquad$ $\chi^2_{1-0.025, 39} = 23.654$

$\chi^2_{0.025, 39} = 58.120$

a) H_0 : $\sigma = 40$ $\quad\quad$ H_1 : $\sigma > 40$

$$\chi^2 = \frac{(n-1)\,S^2}{\sigma^2} = \frac{(40-1)\,(48.5)^2}{(40)^2} = 57.34 > \chi^2_{0.05,39}$$

Hence, the standared deviation is greater than 40.

b) H_0 : $\sigma = 40$ $\quad\quad$ H_1 : $\sigma \neq 40$

We have that $\chi^2_{1-\alpha/2,39} < \chi^2 < \chi^2_{\alpha/2,39}$ so we cannot reject the null hypothesis. It does not significantly differ from 40.

* Two Sample Test :

Let $N(\mu_x, \sigma_x^2)$ and $N(\mu_y, \sigma_y^2)$ be two independent populations $x_1, \dots, x_n$ is a random sample from $N(\mu_x, \sigma_x^2)$ and $y_1, \dots, y_m$ is a random sample from $N(\mu_y, \sigma_y^2)$ $(n \neq m)$

H_0 : $\sigma_x^2 = \sigma_y^2$

H_1 : $\sigma_x^2 < \sigma_y^2$, $\sigma_x^2 > \sigma_y^2$, $\sigma_x^2 \neq \sigma_y^2$

$$\frac{S_x^2(n-1)}{\sigma_x^2} \sim \chi^2_{n-1} \qquad \frac{S_y^2(n-1)}{\sigma_y^2} \sim \chi^2_{m-1}$$

$$\frac{\chi_x^2}{\chi_y^2} = \frac{S_x^2\,(n-1)\,\sigma_y^2}{S_y^2\,(m-1)\,\sigma_x^2} \sim \frac{\chi^2_{n-1}}{\chi^2_{m-1}}$$

Uneler H_0 : $\frac{S_x^2}{S_y^2} = \frac{\chi_x^2/(n-1)}{\chi_y^2/(m-1)} \sim F_{n-1,\,m-1}$

Rejection Region : $\sigma_x^2 > \sigma_y^2$

$F > F_{\alpha, n-1, m-1}$

$\sigma_x^2 < \sigma_y^2$

$F < F_{1-\alpha, n-1, m-1} = \frac{1}{F_{\alpha, n-1, m-1}}$

$\sigma_x^2 \neq \sigma_y^2$

$F > F_{\frac{\alpha}{2}, n-1, m-1}$ or $F < F_{1-\frac{\alpha}{2}, n-1, m-1}$

* A scientist is intenested in whether male and female name driving behaviours. The question she learned was as follows "Is the mean fastest speed driven by a male student different than the mean fastest speed driven by the female students. The psychologist conducted a survey of 34 male students and 29 female students. The statistics she recieved :

	Male	Female
Mean	105.5	90.9
Sd	20.1	12.2

Is there sufficient evidence at $\alpha = 0.05$ level to conclude that the variance of the fastest speed driven by male student differs from the fastest speed driven by female students.

$H_0 : \sigma_1^2 = \sigma_2^2$ $H_1 : \sigma_1^2 \neq \sigma_2^2$ 1 : Male 2 : Female

Under H_0 : $\frac{S_1^2 (n-1)}{S_2^2 (m-1)} = \frac{(20.1)^2}{(12.2)^2} = 2.714 = F$

$F_{0.025, 33, 28} = 2.089$

$F_{1-0.025, 33, 28} = \frac{1}{\quad}$

$F > F_{0.025, 33, 28}$ so we reject the null so the fastest mean differs

2.089

between male and female.

* Question 1 :

Let $X_1, \dots, X_n \overset{iid}{\sim} Ber(p)$ denote n tosses of a baised coin and assume that $\hat{p} = \bar{x}$. We will explore two ways to construct a 95% confidence interval both based on CLT. $\sqrt{n}(\hat{p} - p) \longrightarrow N(0, p(1-p))$

a) Use a plugin estimate $\hat{p}(1-\hat{p})$ for the variance $p(1-p)$ to obtain a 95% CI for p. Use the Wald confidence interval.

By CLT $E(p) = \hat{p}$ and $Var(p) = \hat{p}(1-\hat{p})$

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1) \Rightarrow \mathbb{P}_p\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\hat{p}-p)}{\sqrt{\hat{p}(1-\hat{p})}} \leq z_{\alpha/2}\right) = 0.95$$

$$\Rightarrow p \in \left[\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

b) We can use the CLT which implies for large n

$\mathbb{P}_p\left(-\sqrt{p(1-p)}\, z_{\alpha/2} \leq p \leq \sqrt{p(1-p)}\, z_{\alpha/2}\right) = 1-\alpha$

Solve the equation $\sqrt{n}(\hat{p}-p) = \sqrt{p(1-p)}\, z_{\alpha/2}$ for p in terms of $\hat{p}$ and also solve $\sqrt{n}(\hat{p}-p) = -\sqrt{p(1-p)}\, z_{\alpha/2}$ for p in terms of $\hat{p}$ to obtain a 95 CI for p.

$$n(\hat{p}-p)^2 \leq p(1-p) z^2_{\alpha/2}$$

$$n\hat{p}^2 - 2n\hat{p}p + np^2 \leq p z^2_{\alpha/2} - p^2 z^2_{\alpha/2}$$

$$(n + z^2_{\alpha/2})p^2 - (2n\hat{p} + z^2_{\alpha/2})p + n\hat{p}^2 \leq 0$$

p is between the two real roots of this equation

$$\Delta = (2n\hat{p} + z^2_{\alpha/2}) - 4(n + z^2_{\alpha/2})n\hat{p}^2$$

$$\Rightarrow p \in \left[\frac{(2n\hat{p} + z^2_{\alpha/2}) \pm \sqrt{\Delta}}{2(n + z^2_{\alpha/2})}\right]$$

c) Perform a simulation study to determine the true coverage of the CIs of parts a and b for the combinations:

$n = (10, 40, 100)$

$p = (0.1, 0.3, 0.5)$

For each combination perform at least $B = 100,000$ simulations in each simulation you may simulate $\hat{p} \sim Bin(n, p)$ instead of simulating $x_1, \ldots, x_n$. Report the simulated coverage of the both CIs.

* Question 2:

In a classic genetic study, it was recorded that hospital records gender ratios. The following table shows # of male children out of 6115 families each having 12 children.

# of Male children	# families
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Let $X_1, \dots, X_{6115}$ denote the # of male children.

a) Suggest two reasonable tests for testing $H_0: X_1, \dots, X_{6115} \overset{iid}{\sim} Bin(12, \frac{1}{2})$

$$T_2 = \sum_{i=0}^{12} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=0}^{12} \frac{(O_i - 6115\binom{12}{i}(\frac{1}{2})^{12})^2}{6115\binom{12}{i}(\frac{1}{2})^{12}}$$

$$T_1 = \frac{\text{\# of male} \times \text{frequency}}{\sum \text{frequency}} \quad \text{mean (\# of males)}$$

b) Perform a simulation to simulate the null hypothesis $B = 10^5$. Plot the histograms of T_1 and T_2. Using the simulated values compute the approximate p-values of the tests.

c) In this example, why won't the null hypothesis not work.

No, because there are biological and sociological factors affecting this ratio.

* Question:

Consider the problem of test $X_1, \dots, X_n \sim N(0, 1)$ and H_1: $X_1, \dots, X_n \sim N(\mu, 1)$ where $\mu > 0$. We can use the LRT and the t-test.

a) For $n = 100$ verify numerically that these tests have significance level $\alpha = 0.05$ by performing 10^5 simulations and draw a sample from $N(0, 1)$ computed the test statistics relevant to the LRT (reject H_0: $\bar{x} > \frac{z_{0.05}}{\sqrt{n}}$) and t-test (reject H_0: $\frac{\sqrt{n}\,\bar{x}}{s} > t_{n-1, 0.05}$) and recorded whether each test accepts /

rejects H_0. Record the fraction of simulations that rejects 0.05 and check that it is close to 0.05. Set.seed(1)

Add to them W^+, S. Wilcxon signed rank test rejects H_0 when $W_+ > \frac{n(n+1)}{4} + \sqrt{\frac{n(n+1)(2n+1)}{24}} z_{0.05}$.

For the sign test, reject H_0 if $S > \frac{n}{2} + \sqrt{\frac{n}{4}} z_{0.05}$ where S is the # of +ve values in $x_1, \ldots, x_n$.

b) For $n = 100$, numerically compute the powers of the test against the alternative H_1 for values of $\mu = (0.1 : 0.4)$ do this by performing 10^5 simulations by drawing $X \sim N(\mu, 1)$. Report your findings in a table or visual format.

c) How do the power of the 4 test compare when testing against H_0. You should always use the testing procedure that makes the fews dist. assumptions because in practise we never know whether $\sigma^2 = 1$ or the dist is truely normal. Comment on this statement. John A Rice says it has been shown that the Wilcxon signed test is recently as powerful as t-test thus use it for small n. Do your simulated results support this statement.

* Let's start with one sample setting when $x_1, \ldots, x_n \overset{iid}{\sim} f$, where we drop the normality assumption and we wish to test.

H_0: f is symmetric around 0

H_1: f is symmetric around μ for some $\mu > 0$

Because the shape of f is arbitrary under H_0, the distribution of the t-statistic is no longer the same under every data dist.

P ∈ H₀ in particular, it can be very far from t_{n-1} if n is moderately small and f is heavy-tailed. We consider instead the signed rank statistic W_+ defined in the following way:

1) Sort $|X_1|, |X_2|, \ldots, |X_n|$ in an increasing order. Assign the smallest value a rank of 1, the next 2, and the largest valued a rank n.

2) Define W_+ as the sum of the ranks corresponding to only the +ve values of $X_1, X_2, \ldots, X_n$.

* Example : Wilcoxon Sign Rank Test

X_i : 5, -2, 3, 7

$|X_i|$: 5, 2, 3, 7

Rank : 3, 1, 2, 4

$W_+ = \sum_i$ Rank of $|X_i|$, $X_i > 0$

$= 3 + 2 + 4$

$= 9$

We expect W_+ to be larger under H_1 than under H_0 because higher rank obs are more likely to be +ve under H_1. The test that rejects for large W_+ has the same distribution under every P ∈ H_0 and provides a method for determining the null distribution and rejection threshold for W_+ when n is large. When n is small, we can determine the exact null dist. of W_+ by computing W_+ for all 2^n possible combinations of + and - signs of ranked data.

* **Theorem** : The dist of W_+ is the same for every pdf that is symmetric around 0 for large n, this distribution is asymptotically $N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$ i.e

$$\sqrt{\frac{24}{n(n+1)(2n+1)}}\left(W_+ - \frac{n(n+1)}{4}\right) \xrightarrow{d} N(0,1) \text{ as } n \to +\infty$$

$\frac{n(n+1)(2n+1)}{24}$ $\left(\quad\quad 4\quad\quad\right)$

* **Proof :**

We will show that the distribution of W_+ is the same for every f and that $\mathbb{E}(W_+) = \frac{n(n+1)}{4}$ and $\text{Var}(W_+) = \frac{n(n+1)(2n+1)}{24}$

Let $f_0(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$ be the joint pdf of the data. By symmetry of f about 0, $f(\pm x_1, \ldots, \pm x_n)$ is the same for each of the 2^n combinations of $+/-$ signs. This implies that conditional on $|X_1|, \ldots, |X_n|$, the sign of $X_1, \ldots, X_n$ are independent and each equal to $+$ or $-$ with prob $\frac{1}{2}$.

Let $I_k = 1$, if the value of the rank k is +ve and $I_k = 0$ if the value is -ve. Thus, $I_1, \ldots, I_n \overset{iid}{\sim} \text{Ber}(1/2)$ for any pdf f that is symmetric about zero. The signed rank statistic is $W_+ = \sum_{k=1}^{n} k I_k$ Since $I_1, \ldots, I_n$ have the same distribution under any symmetric pdf about 0, the distribution of W^+ is the same for all such pdfs f.

$$\mathbb{E}(W_+) = \sum_{k=1}^{n} k\, \mathbb{E}(W_+) = \sum_{k=1}^{n} \frac{k}{2} = \frac{1}{2} \cdot \frac{n(n+1)}{2} = \frac{n(n+1)}{4}$$

$$\text{Var}(W_+) = \sum_{k=1}^{n} k^2 \underbrace{\text{Var}(W_+)}_{p(1-p)} = \frac{1}{4} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{n(n+1)(2n+1)}{24}$$

To explain why W_+ is asymptotically normally distributed, we define the empirical cdf of $|X_1|, |X_2|, \ldots, |X_n|$ by :

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{|X_i| \le t\}$$

$F_n(t)$ is the fraction of values of $|X_i|$ that are at most t. The rank of $|X_i|$ can be computed as $n F(|X_i|)$ so $W_+ = \sum_{i=1}^{n} n F_n(|X_i|)\, \mathbb{1}(X_i > 0)$

when n is large, we may show that $F_n(t)$ is with high probability close to the true CDF $F(t)$ of $|X_i|$ for every $t \in \mathbb{R}$, and hence the difference between W_+ and $\tilde{W}_+ = \sum_{i=1}^{n} n F(|X_i|) \mathbb{1}(X_i > 0)$ is negligible. $\tilde{W}_+$ is the sum of iid RVs $Y_i = n F(|X_i|) \mathbb{1}(X_i > 0)$ and hence asymptotically normal by CLT.

By CLT : $$\frac{W_+ - \left(\frac{n(n+1)}{4}\right)}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} N(0,1) \text{ as } n \to +\infty$$

* Rank Sum Test :

The idea of converting observed data values to just rank so as to deal with heavy tailed data and deviations from normality, can be extended to the test sample setting. Consider two indep. samples $X_1, X_2, \dots, X_n \overset{iid}{\sim} f$ and $Y_1, Y_2, \dots, Y_n \overset{iid}{\sim} g$ where f and g are any two arbitrary pdf and we test :

H_0 : $f = g$

H_1 : f stochastically dominates g

In other words, the alternative hypothesis suggests that values drawn from f tend to be larger that values from g. The rank sum test statistic is defined as :

1) Consider the pooled sample of all obs. $X_1, \dots, X_n, Y_1, \dots, Y_m$. Sort them in increasing order and assign the rank 1 to the smallest value 1 and rank $n+m$ for the largest.

2) Define the rank sum test statistic T_Y as the sum of ranks corresponding to only Y_i values i.e values from the second sample.

	x_i	y_i
ex.	2	1
	4	7
	10	8

Values : 1 2 4 7 8 10

Ranks : 1 2 3 4 5 6

T_y : 1 + 4 + 5 = 10

We expect T_y to be smaller under H_1 than under H_0 because H_1 the values of Y_i tend to have smaller ranks. The test that rejects for small values of T_y is called as Wilcoxn rank sum test or alternatively as Mann - Whitney Wilcoxn test. If the alternative H_1 : $f \neq g$ then we would reject for both large & small values of T_y.

The following theorem states that T_y has the same distribution for every $P \in H_0$ and provides a method for determining the null hypothesis and rejection threshold when n and m are both large.

* **Theorem** :

The distribution of T_y is the same under any pdf $f = g$ for large n and m this dist is approximately $N\left(\frac{n(n+m)}{2}, \frac{mn(m+n+1)}{12}\right)$.

If $f = g$, then each ordering of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is equally likely. Since T_y depends only on this ordering, its dist must be the same under every pdf $f = g$.

If $I_k = 1$, if the k^{th} largest value of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ belongs to the second sample and $I_k = 0$ otherwise $I_k = 0$. Then, $T_y = \sum_{k=1}^{m+n} k\, I_k$. Under H_0, I_k indicates whether the k^{th} individual is selected in a simple of random sample of size m (without replacement) from a population of distribution of size m+n.

$$E(T_Y) = \frac{m(m+n+1)}{2}$$

Here the I_k's are not independent

$$Var(T_Y) = \frac{mn(m+n+1)}{12}$$

* Permutation & Randomized Test :

The main idea behind the one-sample Wilcoxon signed rank test & the two-sample ranked sum test is to exploit the symmetry under H_0. For the sign-rank test, the symmetry is that it is equally likely to observe $\pm x_1, \dots, \pm x_n$ for each of the 2^n combinations of $+1$-signs. For the rank-sum test, the symmetry is that it is equally likely to observe each of the $(m+n)!$ permutations of the pooled sample $x_1, \dots, x_n, y_1, \dots, y_m$. In fact, this idea of exploiting the symmetry provides an alternative simulation based method of obtaining the null dist. for any test statistic T.

* Example :

Consider $x_1, \dots, x_n, y_1, \dots, y_m$ and any test statistic depending of x_i's and y_j's ($T = \bar{x} - \bar{y}$ as e.g.). For a null hypothesis H_0 which specifies that all data from both samples are iid from a common dist.

H_0 : $x_1, \dots, x_n, y_1, \dots, y_n \overset{iid}{\sim} f$ for any unknown pdf f.

The permutation null dist. of T is the dist of $T(x_1^*, \dots, x_n^*, y_1^*, \dots, y_n^*)$ when we fix the observed values of $x_1, \dots, x_n, y_1, \dots, y_m$ and let $x_1^*, \dots, y_n^*$ be a permutation of $x_1, \dots, x_n, y_1, \dots, y_m$ chosen

uniformly at random from a set of all $(m+n)!$ possible permutations. Under H_0, each of these $(m+n)!$ possible values of T is equally likely to be observed. To perform a test that rejects for the large values of T, we may use the following procedure:

1) Randomly permute the pooled data $B = 10^4$ times and compute the value of T each time.

2) Compute an approximate p-value as a fraction of B simulations where we observed a value of T larger than T observed, the value for the original (unpermuted) data. Reject at level α if the p-value is at most α.

For a two-sided test that rejects for both large and small values of t, we can compute the p-value by taking the frac of sim where $t > t_{obs}$ or the frac of sim where $t < t_{obs}$ and multiply by 2. This is called a permutation based on t and an example of conditional test because we are looking at the conditional data under H_0 given the set of their values. The utility of this idea may be applied to test statistic t where we don't understand its unconditional distribution under H_0 and where this dist may vary for different pdfs $f = g$.

* Example :

$x_1, \ldots, x_n \in \mathcal{X}$, $y_1, \ldots, y_m \in \mathcal{X}$ be two random samples of obj. represented in some data space $\mathcal{X}$. Suppose you have a f^n

$d(x, y)$ that measures the distance between any two obj. $x, y \in \mathcal{X}$. To test whether $x_1, \ldots, x_n, Y_1, \ldots, Y_m$ appear to come from the same distribution, the following test statistic is used

$$T_1 = \frac{2}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} d(X_i, Y_j) - \frac{1}{\binom{n}{2}} \sum_{1 \le i < i' \le n} d(X_i, X_{i'}) - \frac{1}{\binom{m}{2}} \sum_{1 \le j < j' \le m} d(Y_j, Y_{j'})$$

Hence, T_1 measures whether on average objects from the sample are more similar to each other than different samples.

Or we might consider a "nearest-neighbors" statistic: For each of the $m + n$ data values, look at the k other data values closest to it (as measured by the distance d) and count how many of these come from the same sample as itself. Let T_2 be the average of this count across all $m + n$ data points. So T_2 measures whether the k closest other objects tend to come from the same sample.

The distributions of T_1 and T_2 under H_0 may be difficult to understand theoretically and may depend on the unknown common distribution of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, but we can still carry out a permutation test based on T_1 or on T_2.

A similar idea may be applied in the one-sample setting for testing the null hypothesis

$$H_0 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} f, \text{ for some PDF } f \text{ symmetric about } 0$$

based on the symmetry underlying the Wilcoxon signed-rank test.

* Generalised Likelihood Ratio Test :

$H_0 : \theta = \theta_0$

$H_1 : \theta \neq \theta_0$

$$\Lambda = \frac{\text{lik}(\theta_0)}{\max_{\theta \in \Omega} \text{lik}(\theta)}$$

Ω is the space ; our dist is $\{f(x|\theta) \mid \theta \in \Omega\}$

Hypothesis Testing for categorical data : GLRT

1) Simple H_0 : Let $\{f(x|\theta) \mid \theta \in \Omega\}$ be a parametric model $\theta_0 \in \Omega$ be a particular parameter value. We want to test :

$$H_0 : \quad \theta = \theta_0$$
$$H_1 : \quad \theta \neq \theta_0$$

The GLRT rejects for small values of the test statistic

$$\Lambda = \frac{lik(\theta_0)}{\max_{\theta \in \Omega} lik(\theta)}.$$

The numerator is the value of the likelihood at θ_0 where the denominator is the value of the likelihood at the MLE $\hat{\theta}$.

The level α - test rejects H_0 when $\Lambda \leq c$ where c is chosen; $P_{H_0}(\Lambda \leq c) \approx \alpha$ (probability of Type I error).

Note that the GLRT differs from the LRT in the context of the NP Lemma, where the denominator was instead given by $lik(\theta_1)$ for the simple alternative $\theta = \theta_1$. Since the H_1 is not simple anymore the GLRT replaces the denominator by the maximum value of the likelihood over all values of θ.

* **Problem** : Let $X_1, \dots, X_n \overset{iid}{\sim} N(\theta, 1)$ $H_0 : \theta = 0$, $H_1 : \theta \neq 0$

$$\prod_{i=1}^{n} lik(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_i^2}{2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} x_i^2\right)$$

$$\max lik(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - \bar{x})^2}{2}\right) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} (x_i - \bar{x})^2\right)$$

$$\omega \in \Omega \qquad \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \qquad \alpha \qquad (2\pi)^{n/2} \qquad \alpha \sum_{i=1}^{n}$$

$$\begin{aligned} \Lambda &= \exp\left(\frac{1}{2}\sum_{i=1}^{n}\left((x_i-\bar{x})^2 - x_i^2\right)\right) \\ &= \exp\left(\frac{1}{2}\sum_{i=1}^{n}\left(\bar{x}^2 - 2x_i\bar{x}\right)\right) \\ &= \exp\left(-\frac{n}{2}\bar{x}^2\right) \end{aligned}$$

$\Rightarrow -2\log(\Lambda) = n\bar{x}^2$ ($\Lambda < c$ reject H_0 or $-2\log(\Lambda) > c$ reject)

$\Rightarrow -2\log(\Lambda) = n\bar{x}^2$

Under H_0: $\bar{x} \sim N(0, \frac{1}{n}) \Rightarrow n\bar{x} \sim N(0,1)$

$\Rightarrow n\bar{x}^2 \sim \chi^2_1$

$-2\log(\Lambda) = n\bar{x}^2 \overset{H_0}{\sim} \chi^2_1 \Rightarrow P_{H_0}(-2\log(\Lambda) > K) = \alpha$

$\Rightarrow K = \chi^2_{1,0.05}$ (upper α point)

* In general, the exact sampling distribution of $-2\log(\Lambda)$ under H_0 may not have a simple form as earlier but it may be approximated by a χ^2 dist. for large n.

* **Theorem** :

Let $\{f(x|\theta) \mid \theta \in \Omega\}$ be the parametric model & $X_1, \ldots, X_n \overset{iid}{\sim} f(x|\theta_0)$. Suppose θ_0 is an interior point of Ω and the regularity conditions for consistency & asymptotic normality of the MLE hold. Then, $-2\log\Lambda \to \chi^2_K$ in distribution as $n \to +\infty$ where K = dimension of Ω.

* **Proof** :

We will consider the case where $K = 1$ so θ is the single parameter. Let

$l(\theta)$ be the log-likelihood function and $\hat{\theta}$ be the MLE.

$$\Lambda = \frac{lik(\theta_0)}{lik(\hat{\theta})}$$

Applying a Taylor series expansion of $l(\theta_0)$ around $\theta_0 = \hat{\theta}$

$$\begin{aligned} l(\theta_0) &\approx l(\hat{\theta}) + (\theta_0 - \hat{\theta}) \cancel{l'(\hat{\theta})}^{0} + (\theta_0 - \hat{\theta})^2 \frac{l''(\hat{\theta})}{2} \\ &\approx l(\hat{\theta}) - \frac{1}{2} n I(\theta_0) (\theta_0 - \hat{\theta})^2 \end{aligned}$$

$$\begin{aligned} \Rightarrow -2 \log(\Lambda) &= -2 l(\theta_0) + 2 l(\hat{\theta}) \\ &= n I(\theta_0) (\theta_0 - \hat{\theta})^2 \end{aligned}$$

$\sqrt{n I(\theta_0)} (\hat{\theta} - \theta_0) \longrightarrow N(0,1)$ in distribution by asymptotic normality of MLE

$\Rightarrow \sqrt{n I(\theta_0)} (\hat{\theta} - \theta_0)^2 = -2 \log(\Lambda) \longrightarrow \chi^2_1$ in distribution as $n \to +\infty$.

This theorem implies that an approximate level α test is given by rejecting H_0 when $-2 \log(\Lambda) > \chi^2_k(\alpha)$, the upper α point of χ^2_k dist.

* The dim K of Ω is the # of free parameters in the model - the # of constraints for instance, in the previous example there was only one parameter so the $\dim(\Omega) = 1$. However, for a multinomial model $p_1, \dots, p_K$ the dim = K - 1 as there is a constraint $\sum_{i=1}^{K} p_i = 1$.

2) H_0 composite : GLRT for testing a sub-model

More generally, let $\Omega_0 \subset \Omega$ be a subset of the parameter space Ω corresponding to a lower dimensional submodel for testing

$H_0 : \theta \in \Omega_0$

$H_1 : \theta \notin \Omega_0$, $\Omega_0 \subset \Omega$

The GLRT statistic is defined as $\Lambda = \frac{\max_{\theta \in \Omega_0} lik(\theta)}{\max_{\theta \in \Omega} lik(\theta)}$.

In otherwords, Λ is the ratio of the likelihood evaluated at the MLE is the submodel and at the MLE in the full model.

For large n, under any $\theta_0 \in \Omega_0$, $-2\log(\Lambda)$ is approximately χ^2_K where K is the dimensionality of Ω_0 and Ω and an approximate level α - test rejects when $-2\log(\Lambda) > \chi^2_K(\alpha)$ $K = \dim(\Omega) - \dim(\Omega_0)$.

Example 6.2.1. *(Hardy-Weinberg equilibrium[9]). At a single diallelic locus in the genome with two possible alleles A and a, any individual can have genotype AA, Aa, or aa. If we randomly select n individuals from a population, we may model the numbers of individuals with these genotypes as* $(N_{AA}, N_{Aa}, N_{aa}) \sim \text{Multinomial}\,(n, (p_{AA}, p_{Aa}, p_{aa}))$.

When the alleles A and a are present in the population with proportions θ and $1-\theta$, then under an assumption of random mating, quantitative genetics theory predicts that p_{AA}, p_{Aa}, and p_{aa} should be given by $p_{AA} = \theta^2, p_{Aa} = 2\theta(1-\theta)$, and $p_{aa} = (1-\theta)^2$-this is called the Hardy-Weinberg equilibrium. In practice we do not know θ, but we may still test the null hypothesis that Hardy-Weinberg equilibrium holds for some θ:

$$H_0 : p_{AA} = \theta^2,\ p_{Aa} = 2\theta(1-\theta),\ p_{aa} = (1-\theta)^2 \text{ for some } \theta \in (0,1).$$

This null hypothesis corresponds to a 1-dimensional sub-model (with a single free parameter θ) inside the 2-dimensional multinomial model (specified by general parameters p_{AA}, p_{Aa}, p_{aa} summing to 1). We may test H_0 using the GLRT:
The multinomial likelihood is given by

$(N_{AA}, N_{Aa}, N_{aa}) \sim \text{Multinomial}(n, (p_{AA}, p_{Aa}, p_{aa}))$

$$l(p_{AA}, p_{Aa}, p_{aa}) = \binom{n}{N_{AA}, N_{Aa}, N_{aa}} p_{AA}^{N_{AA}} p_{Aa}^{N_{Aa}} p_{aa}^{N_{aa}}.$$

likelihood of H_0

Letting $\hat{p}_{AA}, \hat{p}_{Aa}, \hat{p}_{aa}$ denote the full-model MLEs and $\hat{p}_{0,AA}, \hat{p}_{0,Aa}, \hat{p}_{0,aa}$ denote the sub-model MLEs, the generalized likelihood ratio is

$$\Lambda = \left(\frac{\hat{p}_{0,AA}}{\hat{p}_{AA}}\right)^{N_{AA}} \left(\frac{\hat{p}_{0,Aa}}{\hat{p}_{Aa}}\right)^{N_{Aa}} \left(\frac{\hat{p}_{0,aa}}{\hat{p}_{aa}}\right)^{N_{aa}},$$

so In order to apply the theorem

$$-2\log\Lambda = 2N_{AA}\log\frac{\hat{p}_{AA}}{\hat{p}_{0,AA}} + 2N_{Aa}\log\frac{\hat{p}_{Aa}}{\hat{p}_{0,Aa}} + 2N_{aa}\log\frac{\hat{p}_{aa}}{\hat{p}_{0,aa}}. \quad (1)$$

↳ # of individuals with this gene

The full-model MLEs are given by $\hat{p}_{AA} = N_{AA}/n, \hat{p}_{Aa} = N_{Aa}/n$, and $\hat{p}_{aa} = N_{aa}/n$, by

Example 1.3.4 from Chapter 2. To find the sub-model MLEs, note that under H_0, the multinomial likelihood as a function of θ is

$\alpha^{N_{aa}} \alpha^{2N_{AA}+N_{Aa}} (1-\alpha)^{2N_{aa}+N_{Aa}}$

$$\begin{aligned}\text{lik}(\theta) &= \binom{n}{N_{AA}, N_{Aa}, N_{aa}} (\theta^2)^{N_{AA}} (2\theta(1-\theta))^{N_{Aa}} \left((1-\theta)^2\right)^{N_{aa}} \\ &= \binom{n}{N_{AA}, N_{Aa}, N_{aa}} 2^{N_{Aa}} \theta^{2N_{AA}+N_{Aa}} (1-\theta)^{N_{Aa}+2N_{aa}}.\end{aligned}$$

Maximizing the likelihood over parameters (p_{AA}, p_{Aa}, p_{aa}) belonging to the sub-model is equivalent to maximizing the above over θ. Differentiating the logarithm of the above likelihood and setting it equal to 0, we obtain the MLE

$$\hat{\theta} = \frac{2N_{AA}+N_{Aa}}{2N_{AA}+2N_{Aa}+2N_{aa}} = \frac{2N_{AA}+N_{Aa}}{2n}$$

for θ, which yields the sub-model MLEs

$$\begin{aligned}\hat{p}_{0,AA} &= \left(\frac{2N_{AA}+N_{Aa}}{2n}\right)^2 \\ \hat{p}_{0,Aa} &= 2\left(\frac{2N_{AA}+N_{Aa}}{2n}\right)\left(\frac{N_{Aa}+2N_{aa}}{2n}\right) \\ \hat{p}_{0,aa} &= \left(\frac{N_{Aa}+2N_{aa}}{2n}\right)^2.\end{aligned}$$

Substituting these expressions into equation (1) yields the formula for $-2\log\Lambda$ in terms of the observed counts N_{AA}, N_{Aa}, N_{aa}. The difference in dimensionality of the two models is $2-1=1$, so an approximate level-α test would reject H_0 when $-2\log\Lambda$ exceeds $\chi^2_1(\alpha)$.

2 : full model
1 : sub model

$-2\log\Lambda > \chi^2_1(\alpha)$ reject.

* Test of Independence :

Gender \ Party	Dem	Rep	Independent	
Male	422	273	381	$P_{1,0}$
Female	299	232	365	$P_{2,0}$
	P_{01} +	P_{02} +	P_{03} =	1

N = 1972

In this example the following table cross-classifies a random sample of 1972 people by gender and by political party identification :

39% of females identified as democrat and 25% as republican while 33% of males identified as democrat and 26% as republican. Is there an evidence of an association btw gender & party identification from which the sample was drawn.

Let N_{ij} be the number of people with gender i who voted for party j : $\sum_{j=1}^{3}\sum_{i=1}^{2} N_{ij} = n$ is the # of observations

Let p_{ij} be the probability of belonging to $(i,j)^{th}$ group. We may count N_{ij} as multinomial distribution with prob p_{ij} ($\sum_i \sum_j p_{ij} = 1$)

We can write the total no of people in group i : $\sum_{j=1}^{3} N_{ij} = N_{i,0}$

The prob : $p_{i,0} = \sum_j p_{ij}$ prob of belonging to gender i

$p_{0,j} = \sum_i p_{ij}$ prob of belonging to party j

Test : $H_0 : p_{ij} = p_{i,0}\, p_{0,j}$ $\forall i,j$ Reject $\Rightarrow$ there is an association

Under H_0, dim(submodel) = 5 - 2 = 3 and the dim(full model) = 6 - 1 = 5. For the GLRT has an approximate null distribution with 5 - 3 = 2 degrees of freedom.

To derive the form :

$$\ell_{SM}(X) = \log\left(\binom{n}{N_1,\dots,N_K}\prod_{i=1}^{K} p_i^{N_i}\right) \qquad (N_1,\dots,N_K) \sim (n(p_1,\dots,p_K))$$

$$= \log\left(\binom{n}{N_1,\dots,N_K}\right) + \sum_{i=1}^{K} N_i \log(p_i)$$

Letting $\hat{p}_{0,i}$ denote the MLEs in the submodel Ω_0 and $\hat{p}_i$ denote the MLEs in the full model.

GLRT : $\Lambda = \prod_{i=1}^{K}\left(\frac{p_{0,i}}{\hat{p}_i}\right)^{N_i} \Rightarrow -2\log(\Lambda) = -2\sum_{i=1}^{K} N_i \log\left(\frac{p_{0,i}}{\hat{p}_i}\right)$

Since $\hat{p}_i = \frac{N_i}{n}$, let $E_i = \hat{p}_{0,i} \cdot n$ denote the expected count for outcome i corresponding to Ω ; $-2\log(\Lambda) = -2\sum_{i=1}^{k} N_i \log\left(\frac{N_i}{E_i}\right)$

Under H_0 : $p_{ij} = p_{i0}\, p_{0j} \quad \forall i, j$

$$\text{lik}(p_{1,0}, p_{2,0}, p_{0,1}, p_{0,2}, p_{0,3}) = \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} \prod_{j=1}^{3} (p_{i0}\, p_{0j})^{N_{ij}}$$

$$= \binom{n}{N_{11}, \ldots, N_{23}} \prod_{i=1}^{2} p_{i,0}^{N_{i,0}} \prod_{j=1}^{3} p_{0,j}^{N_{0,j}}$$

$$\ell(p_{1,0}, \ldots, p_{0,3}) = \log\binom{n}{N_{11}, \ldots, N_{23}} + \sum_{i=1}^{2} N_{i,0} \log(p_{i,0}) + \sum_{j=1}^{3} N_{0,j} \log(p_{0,j})$$

$$+ \lambda\left(\sum_{i=1}^{2} p_{i,0} - 1\right) + \mu\left(\sum_{j=1}^{3} p_{0,j} - 1\right)$$

Select $\lambda = -n$ and $\mu = -n$ $\Rightarrow$ $p_{i,0} = \frac{-N_{i,0}}{\lambda} = \frac{N_{i,0}}{n}$

$$p_{0,j} = \frac{-N_{0,j}}{\mu} = \frac{N_{0,j}}{n}$$

$\hat{p}_{ij} = \left(\frac{N_{i,0}}{n}\right)\left(\frac{N_{0,j}}{n}\right)$ $\Rightarrow$ $E_{ij} = n\,\hat{p}_{ij}$

$\Rightarrow$ $-2\log\Lambda = 8.31$

p-value = 0.016 Reject !

Hence, there is an association between gender & political identification

* Test of Homogeneity :

We have two independent count obs from a multinomial samples and each have K outcomes $(N_1, \ldots, N_K) \sim MN(n, (p_1, \ldots, p_K))$ and $(M_1, \ldots, M_K) \sim MN(m, (q_1, \ldots, q_K))$ where n and m are known sample sizes. To test the homogeneity H_0 $p_i = q_i$ $\forall i \in \{1, \ldots, K\}$

The following table counts the occurrences of six different short words in Chapter 1 and Chapter 6 of Sense and Sensibility written by Austen and in Chapters 12 and 24 of Sanditon written by the admirer.

	a	an	this	that	with	without	
Ch 1 & 6	101	11	15	37	28	10	202
Ch 12 & 24	83	29	15	22	43	4	196

Let us model the counts from Ch 1 & Ch 6 $\sim MN(202, (p_1, \dots, p_6))$ and Ch 12 & Ch 24 $\sim MN(196, (q_1, \dots, q_6))$. We wish to test $H_0: p_i = q_i \ \forall i \in \{1, \dots, 6\}$

To derive the GLRT:

$$\text{lik}(p_1, \dots, p_K, q_1, \dots, q_K) = \binom{n}{N_1, \dots, N_K} \prod_{i=1}^{K} p_i^{N_i} \binom{m}{M_1, \dots, M_K} \prod_{i=1}^{K} q_i^{M_i}$$

Let $\hat{p}_i = \frac{N_i}{n}$, $\hat{q}_i = \frac{M_i}{m}$ be the full model MLEs & $\hat{p}_{0,i} = \hat{q}_{0,i}$ be the MLEs of the sub model.

$$\Lambda = \prod_{i=1}^{K} \left(\frac{p_{0,i}}{\hat{p}_i}\right)^{N_i} \prod_{i=1}^{K} \left(\frac{q_{0,i}}{\hat{q}_i}\right)^{M_i} \Rightarrow$$

$$-2\log(\Lambda) = 2 \sum_{i=1}^{K} \left(N_i \log\left(\frac{\hat{p}_i}{p_{0,i}}\right) + M_i \log\left(\frac{\hat{q}_i}{\hat{q}_{0,i}}\right)\right)$$

$$= 2 \sum_{i=1}^{K} \left(N_i \log\left(\frac{\hat{p}_i}{E_i}\right) + M_i \log\left(\frac{\hat{q}_i}{F_i}\right)\right)$$

$$E_i = n \hat{p}_{0,i} \qquad F_i = m \hat{q}_{0,i}$$

In the sub model: $\text{lik}(p_1, \dots, p_K) = \underbrace{\binom{n}{N_1, \dots, N_K}}_{\alpha} \underbrace{\binom{m}{M_1, \dots, M_K}}_{\alpha'} \prod_{i=1}^{K} p_i^{N_i + M_i}$

$$\log \text{lik}(p_1, \dots, p_K) = \log(\alpha) + \log(\alpha') + \sum_{i=1}^{K} (N_i + M_i) \log p_i + \lambda \left(\sum_{i=1}^{K} p_i - 1\right)$$

$$\Rightarrow p_i = \frac{M_i + N_i}{-\lambda} \quad \text{choose} \quad \lambda = -(n+m)$$